Introduction to Bayesian Statistics [1]

Notes

October 15, 2015

1 Introduction to Statistical Science

- Association between variables may be (partly) due to lurking variables, not necessarily causal
- Frequentist (classical) approach
  - Parameters are unknown, but fixed
  - Probabilities are long-term relative frequencies
- Bayesian approach
  - Parameters are random variables
  - Probabilities of parameters are degrees of belief
  - Data revises our prior beliefs → posterior distribution

2 Scientific Data Gathering

- Some definitions
  - *Population*: the whole group of objects under investigation
  - *Sample*: a subset of population for the experiments
  - *Parameter*: characteristics of the population
  - *Statistic*: characteristics of the sample → infer parameters
- Sampling methods
  - Simple: random
  - Stratified: population is divided into strata, samples sizes are proportional to stratum sizes
  - Cluster: localized sampling (cost-effective, but may be biased)
- Errors come from sampling, but also because of human reasons
- Experiment design
  - Completely randomized
  - Randomized block: divide into blocks by an identified variable, and run experiments in each block

3 Displaying and Summarizing Data

- Displaying a single variable
  - *Dotplot*: dots over the number line (equal values stacked vertically)
  - *Boxplot*: line from Q_0 to Q_1, rectangle from Q_1 to Q_2, rectangle from Q_2 to Q_3, and line from Q_3 to Q_4, where Q_i $(i = 0 \dots 4)$ are quartiles of the data, i.e., percentiles of 0%, 25%, 50%, 75%, and 100%, respectively
  - *Stem-and-leaf diagram*: a simple histogram, with bins containing a power of 10; draw a vertical line, the labels of the bins are written on the left, and contain only the digits common to their contents; the next digits of the data are written on the right
  - *Frequency table / Histogram*: using arbitrary bins, but bars of the histogram should have correct relative areas, so a bin with double width should have half the height
  - *Cumulative frequency polygon*: cumulative frequencies of the bins plotted as a curve, easy to see percentiles graphically
- Location measures
  - *Mean*: $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$
    easy to calculate, good properties, but influenced by outliers
  - *Median*: Q_2, the middle of the ordered data (or the avarage of the two middle values when n is even); not influenced by outliers
  - *Trimmed mean*: $\bar{y}_k = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} y_i$
- Spread measures
  - *Range*: $Q_4 - Q_0$; influenced by outliers
  - *Interquantile range* (*IQR*): $Q_3 - Q_1$; does not combine well
  - *Variance*: $\mathrm{Var}(y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$; not comparable to mean (squared)
  - *Standard deviation*: $sd(y) = \sqrt{\mathrm{Var}(y)}$
- Measures of grouped data
  (J: # of bins, m_j: midpoint, n_j: # of data points, R_j: width)
  - *Mean*: $\bar{y} = \frac{1}{n} \sum_{j=1}^{J} n_j m_j$
  - *Variance* (if points in a bin are treated as the midpoint):
    $\mathrm{Var}(y) = \frac{1}{n} \sum_{j=1}^{J} n_j (m_j - \bar{y})^2$
  - *Variance* (if points in a bin are spread out evenly):
    $\mathrm{Var}(y) = \frac{1}{n} \sum_{j=1}^{J} n_j \left[(m_j - \bar{y})^2 + \frac{1}{12} R_j^2 \right]$

- Displaying two or more variables
  - *Scatterplot*: Dotplot with two axes
  - *Scatterplot matrix*: a matrix of scatterplots, where the (i, i) element shows the label of the i-th variable, other (i, j) elements have the i-th variable as the x axis, and the j-th variable as the y axis
- Measures of association
  - *Covariance*: $\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
  - *Correlation*: $\mathrm{Corr}(x, y) = \mathrm{Cov}(x, y)/\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}$ always in $[-1, 1]$, negative values show reverse correlation

4 Logic, Probability, and Uncertainty

- Some definitions
  - *Sample space* (Ω/U): all possible outcomes of a random experiment
  - *Event*: any set of possible outcomes of a random experiment
  - *Mutually exclusive / disjoint events*: no outcomes in common
  - *Union* ($\cup$), *intersection* ($\cap$), *complement* ($\tilde{A}$)
  - *Independent events*: events that do not affect each other, $P(A \cap B) = P(A) \cdot P(B)$
  - *Marginal probability*: probability of one event in a joint setting, e.g. $P(A) = P(A \cap B) + P(A \cap \tilde{B})$ for two events
- Probability axioms
  - $P(A) \geq 0$ for any event A
  - $P(U) = 1$
  - $P(A \cup B) = P(A) + P(B)$ for mutually exclusive events A and B
- Other rules
  - $P(\emptyset) = 0$ [empty set probability]
  - $P(\tilde{A}) = 1 - P(A)$ [complement probability]
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ [addition rule]
  - $P(A) = \sum_{j=1}^{n} P(A \cap B_j)$ [law of total probability]
- Conditional probability
  - A occurred, B is an unobservable event
  - $P(B|A) = P(A \cap B)/P(A)$ [for independent events: $P(B)$]
  - $P(A \cap B) = P(B) \cdot P(A|B)$ [multiplication rule]

- Bayes' theorem
  - $P(B|A) = P(A|B) \cdot P(B) / \left[P(A|B) \cdot P(B) + P(A|\tilde{B}) \cdot P(\tilde{B}) \right]$
  - $P(B)$ is the *prior probability* of B (our belief)
  - $P(A|B)$ is the *likelihood* of event A if B is true
  - $P(B|A)$ is the *posterior probability* of B
  - Only relative likelihood counts (we can omit constant multipliers)
  - For a set of events B_j partitioning the universe:

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{\sum_{j=1}^{n} P(A|B_j) \cdot P(B_j)}$$

- Bayes' theorem for odds
  - $\text{odds}_{\text{prior}}(C) = P(C)/P(\tilde{C})$, as in "the chance is ten to one"
  - $\text{odds}_{\text{post}}(\text{C}) = \text{odds}_{\text{prior}} \cdot B$, where B is the Bayes factor
  - $B = P(D|C)/P(D|\tilde{C})$

5 Discrete Random Variables

- Discrete random variable Y
  - The outcome of a random experiment
  - Can take only separated values y_k
  - *Probability distribution*: $f(y_k) = P(Y = y_k)$
  - *Expected value*: $E(Y) = \sum_k y_k \cdot f(y_k)$
  - *Variance*: $\text{Var}(Y) = E(Y - E(Y))^2 = E(Y^2) - E(Y)^2$
- Linear function measures
  - $E(aY + b) = aE(Y) + b$
  - $\text{Var}(aY + b) = a^2\text{Var}(Y)$
- Binomial distribution
  - $Y = \text{binomial}(n, \pi)$
  - n independent trials
  - π is the probability of success in each trial
  - Y is the number of successes in n trials
  - $f(y|\pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$
  - $E(Y|\pi) = n\pi$
  - $\text{Var}(Y|\pi) = n\pi(1 - \pi)$

- Hypergeometric distribution
  - $Y = \text{hypergeometric}(N, R, n)$
  - Taking n balls from an urn with N balls, R of which are red
  - Y is the number of red balls taken
  - $f(y|N, R, n) = \binom{R}{y} \cdot \binom{N-R}{n-y} / \binom{N}{n}$
  - $E(Y|N, R, n) = n \cdot \frac{R}{N}$
  - $\text{Var}(Y|N, R, n) = n \cdot \frac{R}{N} \cdot (1 - \frac{R}{N}) \cdot \frac{N-n}{N-1}$
- Poisson distribution
  - $Y = \text{Poisson}(\mu)$
  - Y is the # of times a rare event occurs during a period of time
  - μ is the average number of occurrences
  - $f(y|\mu) = \mu^y e^{-\mu}/y!$
  - $E(Y|\mu) = \mu$
  - $\text{Var}(Y|\mu) = \mu$
- Joint random variables
  - $f(x_i, y_j) = P(X = x_i, Y = y_j)$
  - *Independent variables*: $f(x_i, y_j) = f(x_i) \cdot f(y_j)$ for all i, j
  - *Marginal distribution*: $f(y_j) = \sum_i f(x_i, y_j)$
  - $E(h(X, Y)) = \sum_i \sum_j h(x_i, y_j) \cdot f(x_i, y_j)$
  - $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$ for independent variables
- Conditional probability for random variables
  - $Y = y_j$ occurred, X is an unobservable parameter
  - $f(x_i|y_j) = f(x_i, y_j)/f(y_j)$
  - $f(x_i, y_j) = f(x_i) \cdot f(y_j|x_i)$ [multiplication rule]

6 Bayesian Inference for Discrete Random Variables

- Bayes' theorem
  - $g(x_i|y_j) = g(x_i) \cdot f(y_j|x_i) / \left[\sum_k g(x_k) \cdot f(y_j|x_k)\right]$
  - $g(x_i)$ is the *prior probability* function
  - $f(y_j|x_i)$ is the *likelihood* function
  - $g(x_i|y_j)$ is the *posterior probability* function
  - Only relative likelihood counts (we can omit constant multipliers)

- Bayes' theorem for binomial (Poisson) with discrete prior
  - Set up a table with each possible value of π (μ) in the first columns
  - Put the (relative) prior probabilities in the second column
  - Put the (relative) likelihoods in the third column
    - binomial: $\pi_i^{y_j}(1-\pi_i)^{y_j}$ [no binomial multiplier]
    - Poisson: $\mu_i^{y_j}e^{-\mu_i}$ [no factorial divisor]
  - Compute prior times likelihood in the fourth column
  - Compute the marginal sum of the fourth column
  - Put the normalized values into the last (posterior) column

7 Continuous Random Variables

- Measures of continuous random variables
  - *Probability density function*: $f(y)$
  - $\int_{-\infty}^{\infty} f(y)\,\mathrm{d}y = 1$
  - $P(a < y < b) = \int_a^b f(y)\,\mathrm{d}y$
  - $E(Y) = \int_{-\infty}^{\infty} yf(y)\,\mathrm{d}y$
  - $\mathrm{Var}(Y) = E(Y - E(Y))^2 = E(Y^2) - E(Y)^2$
- Uniform distribution = $\beta(1,1)$ [see below]
- β-distribution
  - $X = \beta(a,b)$
  - $g(x|a,b) = k \cdot x^{a-1}(1-x)^{b-1}$ for $0 \le x \le 1$; 0 elsewhere
  - $k = \Gamma(a+b)/\left[\Gamma(a)\Gamma(b)\right]$ normalizing constant
  - $E(X) = a/(a+b)$
  - $\mathrm{Var}(X) = ab/\left[(a+b)^2(a+b+1)\right]$
  - Can be approximated by the normal distribution, when a and b are larger than 10
- γ-distribution
  - $X = \gamma(r,v)$
  - $g(x|r,v) = k \cdot x^{r-1}e^{-vx}$ for $0 \le x < \infty$
  - $k = v^r/\Gamma(r)$ normalizing constant
  - $E(X) = r/v$
  - $\mathrm{Var}(X) = r/v^2$

- Normal distribution
  - $X = \text{normal}(\mu, \sigma^2)$
  - $g(x|\mu, \sigma^2) = k \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
  - $k = 1/(\sqrt{2\pi} \cdot \sigma)$ normalizing constant
  - $E(X) = \mu$
  - $\text{Var}(X) = \sigma^2$
- Central limit theorem
  - Taking a random sample of n elements from an arbitrary distribution with mean μ and variance σ^2, the limiting distribution of $(\bar{y} - \mu)/(\sigma/\sqrt{n})$ is normal$(0, 1)$
  - This is true for relatively small samples, $n \geq 25$ is sufficient
- Marginal density of continuous random variables
  - $f(y) = \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x$, when X is continuous
  - $f(y) = \sum_i f(x_i, y)$, when X is discrete
- Conditional probability density
  - $Y = y$ occurred, X is an unobservable parameter
  - $f(x|y) = f(x, y)/f(y)$

8 Bayesian Inference for Binomial Proportion

- Posterior distribution: $g(\pi|y) = g(\pi) \cdot f(y|\pi)/k$
- Normalizing constant is $k = \int_0^1 g(\pi) \cdot f(y|\pi)\,\mathrm{d}\pi$
- Needs numerical integration unless the prior is special
- Uniform prior: β prior with $a = b = 1$
- β prior
  - $g(\pi|a, b) = \beta(a, b)$
  - $g(\pi|y) = \beta(y + a, n - y + b)$
  - *Equivalent sample size*: $n_{\text{eq}} = a + b + 1$
  - *Mode* (most probable value): $(a - 1)/(a + b - 2)$
  - This is the conjugate prior for the binomial distribution
- Jeffrey's prior: β prior with $a = b = \frac{1}{2}$
  - Invariant under continuous transformation

- Choosing a prior
  - Uniform, when we have no information
  - β by the mean and spread—but check the equivalent sample size, i.e., how much information is in it (should be way less than the data)
  - General prior (needs numerical integration)
- Credible interval
  - A β posterior can be approximated by normal distribution with posterior mean (m) and variance (s)
  - For $(1-\alpha)\cdot 100\%$ credible region: $\pi \approx m \pm z_{\alpha/2}\cdot s$
  - $\int_{-\infty}^{-z_{\alpha/2}} g(x)\,\mathrm{d}x = \int_{z_{\alpha/2}}^{\infty} g(x)\,\mathrm{d}x = \frac{\alpha}{2}$, where g is normal$(0,1)$

9 Comparing Bayesian and Frequentist Inferences for Proportion

- Point estimation: [F] unbiased / [B] biased
- Confidence interval: [F] pre-data / [B] post-data (credible interval)
- Hypothesis tests (one-sided)
  - Bayesian version: reject the hypothesis $H_0:\ \pi \le \pi_0$, when the posterior probability is less than the level of significance, e.g. $P(\pi \le \pi_0|y) < 0.05$ (for 95% significance)
  - Normal distribution constants for one-sided hypothesis tests: $z_{0.01} = 2.327$ (99%), $z_{0.05} = 1.645$ (95%), $z_{0.1} = 1.282$ (90%)
- Hypothesis test (two-sided)
  - Bayesian version: reject the hypothesis $H_0:\ \pi = \pi_0$, when the null π_0 is not in the credible interval
  - Normal distribution constants for credible interval computation: $z_{0.005} = 2.575$ (99%), $z_{0.025} = 1.96$ (95%), $z_{0.05} = 1.645$ (90%)

10 Bayesian Inference for Poisson

- Posterior distribution: $g(\mu|y_1,\ldots,y_n) = g(\mu)\cdot f(y_1,\ldots,y_n|\mu)/k$
- Normalizing constant is $k = \int_0^{\infty} g(\mu)\cdot f(y_1,\ldots,y_n|\pi)\,\mathrm{d}\mu$
- Needs numerical integration unless the prior is special
- Uniform prior: γ prior with $r=1$, $v=0$ [formally]
- Jeffrey's prior: γ prior with $r=\frac{1}{2}$, $v=0$ [formally]
  - Invariant under continuous transformation

- γ prior
  - $g(\mu|r,v) = \gamma(r,v)$
  - $g(\mu|y) = \gamma(r+y, v+1)$,
    so after n samples $g(\mu|y_1,\ldots,y_n) = \gamma(r+\sum_{i=1}^n y_i, v+n)$
  - *Equivalent sample size*: $n_{\text{eq}} = v$
  - *Mode* (most probable value): $(r-1)/v$
  - This is the conjugate prior for the Poisson distribution

11 Bayesian Inference for Normal Mean

- We assume that the variance σ^2 is known
- Discrete prior
  - Done as in Sec. 6
  - The likelihood of a random sample $y_1,\ldots,y_n$ is
    proportional to the likelihood of $\bar{y}$ with mean μ and variance σ^2/n,
    so likelihood $\propto e^{-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2}$
- Flat (Jeffrey's) prior
  - $g(\mu) \propto 1$
  - $g(\mu|y) = \text{normal}(y, \sigma^2)$
  - $g(\mu|y_1,\ldots,y_n) = \text{normal}(\bar{y}, \sigma^2/n)$
- Normal prior
  - $g(\mu) = \text{normal}(m, s^2)$, plausible values are in the $m \pm 3s$ interval
  - $g(\mu|y) = \text{normal}(m_1, s_1^2)$, where $\frac{1}{s_1^2} = \frac{1}{s^2} + \frac{1}{\sigma^2}$ and
    $m_1 = (\sigma^2 m + s^2 y)/(\sigma^2 + s^2) = s_1^2(\frac{1}{s^2}m + \frac{1}{\sigma^2}y)$
  - $g(\mu|y_1,\ldots,y_n) = \text{normal}(m_n, s_n^2)$, where $\frac{1}{s_n^2} = \frac{1}{s^2} + \frac{n}{\sigma^2}$ and
    $m_n = (\frac{\sigma^2}{n}m + s^2\bar{y})/(\frac{\sigma^2}{n} + s^2) = s_n^2(\frac{1}{s^2}m + \frac{n}{\sigma^2}\bar{y})$
  - *Equivalent sample size*: $n_{\text{eq}} = \sigma^2/s^2$
  - *Mode* (most probable value): μ
  - This is the conjugate prior for the normal distribution
- When the variance is unknown
  - *Sample variance*: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$
  - But then the credible interval should be computed
    using the Student's t-distribution with $n-1$ DoF
- Predictive density: $f(y_{n+1}|y_1,\ldots,y_n) = \text{normal}(m_n, \sigma^2 + s_n^2)$

12 Comparing Bayesian and Frequentist Inferences for Mean

- See Sec. 9

13 Bayesian Inference for Difference Between Means

- Independent samples
  - Using the subtraction rules for mean and variance (Sec. 5)
  - When the (equal) variances are unknown use the approximation $\hat{\sigma}^2 = \frac{1}{n_1+n_2-2}\left[\sum_i (y_{i1}-\bar{y}_1)^2 + \sum_j (y_{j2}-\bar{y}_2)^2\right]$, but then use the Student's t-distribution with n_1+n_2-2 DoF
  - When the unknown differences are different, use $\hat{\sigma}_1^2 = \frac{1}{n_1-1}\sum_i (y_{i1}-\bar{y}_1)^2$ and $\hat{\sigma}_2^2 = \frac{1}{n_2-1}\sum_i (y_{i2}-\bar{y}_2)^2$, and Student's t-distribution with this DoF (rounded):
  $$\frac{\left(\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}\right)}{\frac{(\hat{\sigma}_1^2/n_1)^2}{n_1+1}+\frac{(\hat{\sigma}_2^2/n_2)^2}{n_2+1}}$$
  - For difference of proportions, use a normal approximation of the posteriors with the same mean and variance
- Paired experiments (dependent samples)
  - This is a two-element randomized block design (Sec. 2)
  - Use the differences as a single sample

14 Bayesian Inference for Simple Linear Regression

- Least squares fit
  - x is the *predictor*, y the *response* variable
  - We search for a line $y = \alpha_{\bar{x}} + \beta x$, where $\alpha_{\bar{x}}$ is the intercept at $\bar{x}$, and β is the slope (any intercept is OK, but $\bar{x}$ is useful)
  - $\hat{y} = \bar{y} + B(x-\bar{x})$, where $B = (\overline{xy} - \bar{x}\bar{y})/(\overline{x^2} - \bar{x}^2)$
  - $y_i - \hat{y}_i$ is the *residual* (error)
  - Variance of the points around the LSQ line: $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$, because we lost 2 DoF
  - If the points' relation is exponential, we can do a LSQ fit on the logarithms of the data

- Linear regression assumptions
  - The mean of y given x is a linear function of x
  - The error's distribution is normal$(0, \sigma^2)$, where the variance is known and equal for all y
  - The errors are independent of each other
- Bayesian inference
  - likelihood$(\alpha_{\bar{x}}, \beta) = \underbrace{\text{normal}(\bar{y}, \sigma^2/n)}_{\text{likelihood}(\alpha_{\bar{x}})} \cdot \underbrace{\text{normal}(B, \sigma^2/SS_x)}_{\text{likelihood}(\beta)}$, where $SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2 = n(\overline{x^2} - \bar{x}^2)$ is the sum of squares
  - Priors: $g(\alpha_{\bar{x}}, \beta) = g(\alpha_{\bar{x}})g(\beta)$ [both normal or flat]
  - When σ^2 is unknown, use the estimated variance $\hat{\sigma}^2$ instead, but use the Student's t with $n-2$ DoF for credible interval / hypotheses
  - Predictive distribution: $f(y_{n+1}|x_{n+1}) = \text{normal}(\hat{m}_{n+1}, \hat{s}^2_{n+1} + \sigma^2)$
  - $\hat{m}_{n+1}$ and $\hat{s}^2_{n+1}$ are the mean and variance of the LSQ line at x_{n+1}: $\hat{m}_{n+1} = m_{\alpha_{\bar{x}}} + m_\beta \cdot (x_{n+1} - \bar{x})$, $\hat{s}^2_{n+1} = s^2_{\alpha_{\bar{x}}} + s^2_\beta \cdot (x_{n+1} - \bar{x})^2$, where $m_{\alpha_{\bar{x}}}$, $s^2_{\alpha_{\bar{x}}}$, m_β, and s^2_β are posterior parameters

15 Bayesian Inference for Standard Deviation

- $S \times \chi^{-2}$-distribution ("S times inverse chi-squared")
  - $X = S \times \chi^{-2}(\kappa)$, where κ is the DoF
  - $g(x|S, \kappa) = k \cdot x^{-\frac{\kappa}{2}-1} e^{-\frac{S}{2x}}$
  - $k = S^{\frac{\kappa}{2}} / \left[2^{\frac{\kappa}{2}}\Gamma(\kappa/2)\right]$
  - $E(X) = S/(\kappa - 2)$, when $\kappa > 2$
  - $\text{Var}(X) = \frac{2S^2}{(\kappa-2)^2(\kappa-4)}$, when $\kappa > 4$
  - *Mode* (most probable value): $S/(\kappa + 2)$
  - S/X has χ^2-distribution with κ DoF
- Inference for σ^2 and σ
  - likelihood $\propto SS_T \times \chi^{-2}(n-2)$, where $SS_T = \sum_i (y_i - \mu)^2$
  - When μ is unknown, use $\bar{y}$ and deduct one DoF: likelihood $\propto SS_y \times \chi^{-2}(n-3)$, where $SS_y = \sum_i (y_i - \bar{y})^2$
  - Change of variable between σ^2 and σ: $g_\sigma(\sigma) = g_{\sigma^2}(\sigma^2) \cdot 2\sigma$
- Uniform prior for σ^2 ($= 0 \times \chi^{-2}$ prior with $\kappa = -2$ [formally])
  - $g_{\sigma^2}(\sigma^2) \propto 1$, $g_\sigma(\sigma) \propto \sigma$
  - $g_{\sigma^2}(\sigma^2|y_1, \ldots, y_n) = SS_T \times \chi^{-2}(n-2)$
- Uniform prior for σ ($= 0 \times \chi^{-2}$ prior with $\kappa = -1$ [formally])
  - $g_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma}$, $g_\sigma(\sigma) \propto 1$
  - $g_{\sigma^2}(\sigma^2|y_1, \ldots, y_n) = SS_T \times \chi^{-2}(n-1)$

- Jeffrey's prior ($= 0 \times \chi^{-2}$ prior with $\kappa = 0$ [formally])
  - $g_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma^2}$, $g_\sigma(\sigma) \propto \frac{1}{\sigma}$
  - $g_{\sigma^2}(\sigma^2|y_1, \ldots, y_n) = SS_T \times \chi^{-2}(n)$
- $S \times \chi^{-2}$ prior
  - $g_{\sigma^2}(\sigma^2) = S \times \chi^{-2}(\kappa)$, $g_\sigma(\sigma) \propto (\sigma^2)^{-\frac{\kappa-1}{2}-1} e^{-\frac{S}{2\sigma^2}}$
  - $g_{\sigma^2}(\sigma^2|y_1, \ldots, y_n) = (S + SS_T) \times \chi^{-2}(\kappa + n)$
  - When choosing an $S \times \chi^{-2}$ prior, decide on a median c, then find the S, where $P(\frac{S}{\sigma^2} < \frac{S}{c^2}) = 50\%$, here $\frac{S}{\sigma^2}$ is $\chi^2(\kappa)$ [use $\kappa = 1$ for maximum spread]

16 Robust Bayesian Methods

- Precise, but misspecified prior $\rightarrow$ wrong posterior
- I: misspecification indicator (0: OK, 1: misspecified)
- $g(\theta|i) = \begin{cases} g_0(\theta) & \text{if } i = 0 \text{ [the precise prior]} \\ g_1(\theta) & \text{if } i = 1 \text{ [a flat or vague conjugate prior]} \end{cases}$
- $p_i = P(I = i)$, our confidence in the specific prior (e.g. $p_0 = 95\%$)
- Marginal posterior probability of misspecification:

$$P(I = i|y_1, \ldots, y_n) = \frac{p_i f_i(y_1, \ldots, y_n)}{p_0 f_0(y_1, \ldots, y_n) + p_1 f_1(y_1, \ldots, y_n)},$$

  where f_i is the marginal probability of the data when $I = i$:

$$f_i(y_1, \ldots, y_n) = \int g_i(\theta) \cdot f(y_1, \ldots, y_n|\theta) \, \mathrm{d}\theta$$

- $g(\theta|y_1, \ldots, y_n) = \sum_{i=0}^{1} g_i(\theta|y_1, \ldots, y_n) \cdot P(I = i|y_1, \ldots, y_n)$

Appendix: Student's t-distribution

- $X = t(k)$, where $k > 0$ is the DoF
- $g(x|k) = k \cdot (1 + x^2/k)^{-\frac{k+1}{2}}$
- $k = \Gamma(\frac{k+1}{2}) / \left[\sqrt{k\pi} \cdot \Gamma(k/2)\right]$ normalizing constant
- $E(X) = 0$ for $k > 1$
- $\mathrm{Var}(X) = k/(k-2)$ for $k > 2$, ∞ for $1 < k \leq 2$

References

[1] W. M. Bolstad, *Introduction to Bayesian statistics*. Wiley & Sons, 2013.